Module 19 Multivariate Analysis for Genetic data
Session 10: Multivariate inference

Jan Graffelman[1,2]

[1]Department of Statistics and Operations Research
Universitat Politècnica de Catalunya
Barcelona, Spain

[2]Department of Biostatistics
University of Washington
Seattle, WA, USA

28th Summer Institute in Statistical Genetics (SISG 2023)

Contents

Multivariate normal distribution

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\mu)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mu)}$$

Parameters:

- Population mean vector:

$$\mu = (\mu_1, \mu_2, \ldots, \mu_p)$$

- Population variance-covariance matrix:

$$Cov(\mathbf{X}) = \boldsymbol{\Sigma}_{p\times p} = E\left((\mathbf{X}-\mu)(\mathbf{X}-\mu)'\right) = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix}$$

Inference on a mean vector

Univariate test on a population mean

- $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$
- Statistic: $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} \sim t_{n-1}$
- $100 \cdot (1-\alpha)$ confidence interval: $CI_{1-\alpha}(\mu) = \overline{x} \pm t_{n-1,\alpha/2} s/\sqrt{n}$

Note that

$$t^2 = \frac{(\bar{x}-\mu_0)^2}{s^2/n} = n(\overline{x}-\mu_0)(s^2)^{-1}(\overline{x}-\mu_0)$$

By analogy, for the multivariate case we obtain Hotelling's T^2

$$T^2 = n(\overline{\mathbf{x}}-\mu_0)'\mathbf{S}^{-1}(\overline{\mathbf{x}}-\mu_0)$$

Multivariate test on a population mean vector

- $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$
- Statistic: $\frac{(n-p)}{p(n-1)} T^2 \sim F_{p,n-p}$
- $100 \cdot (1-\alpha)$ confidence region is the ellipse traced for μ:

$$n(\overline{\mathbf{x}}-\mu)'\mathbf{S}^{-1}(\overline{\mathbf{x}}-\mu) \leq c^2 = \frac{(n-1)p}{n-p} F_{p,n-p}(\alpha)$$

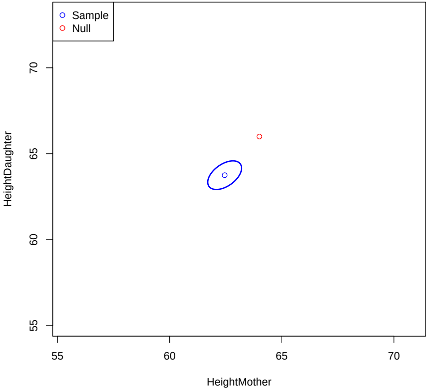
Example

Height of mothers and daughters of the Pearson-Lee data (1903)

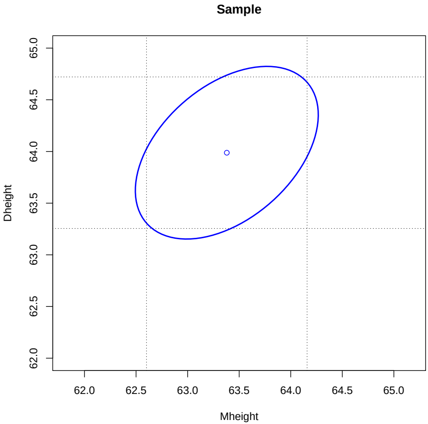
$$H_0 : (\mu_M, \mu_D) = (64, 66) \text{ vs } H_0 : (\mu_M, \mu_D) \neq (64, 66)$$

$$T^2 = 562.9, \text{ df1} = 2, \text{ df2} = 1373, \text{ p-value} < 2.2\text{e-}16$$

Confidence region



Confidence region versus confidence intervals



Univariate Student t-test for two independent samples (common σ^2)

Hypothesis:

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases}$$

Test statistic:

$$T = \frac{\overline{x}_m - \overline{x}_n - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{m} + \frac{1}{n}}}$$

$$s_p^2 = \frac{(m-1)\, s_X^2 + (n-1)\, s_Y^2}{n + m - 2}$$

Under the null:

$$T \sim t_{n+m-2}$$

Example

	N	N*	Mean	Stdev	Med	Q1	Q3	Min	Max
Boys	77	0	179.506	6.5	178	175	183	165	198
Girls	14	0	167.5	4.363	168.5	165	170	160	174

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases}$$

$$s_p^2 = \frac{(m-1)\,S_X^2 + (n-1)\,S_Y^2}{n+m-2} = \frac{(77-1)(6.5)^2 + (14-1)(4.363)^2}{77+14-2} = 38.86232$$

$$T = \frac{\overline{X}_m - \overline{Y}_n - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{m} + \frac{1}{n}}} = \frac{179.506 - 167.5}{\sqrt{38.86232}\sqrt{\frac{1}{77} + \frac{1}{14}}} = 6.628885$$

Critical value: $t_{89,0.975} = 1.986979$ p-value: $2 \cdot P\left(t_{89} > 6.628885\right) = 2.52e - 09$

$$CI_{0.95}(\mu_1 - \mu_2) = \left(\left(\overline{X}_m - \overline{Y}_n\right) \pm t_{n+m-2,\alpha/2}\, s_p \sqrt{\frac{1}{m} + \frac{1}{n}} \right) = (8.408, 15.605)$$

Multivariate comparison of two groups (common $\boldsymbol{\Sigma}$)

$$H_0 : \mu_1 = \mu_2 \text{ vs } H_1 : \mu_1 \neq \mu_2$$

Assumptions:

- Both populations are multivariate normal
- $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$

Results:

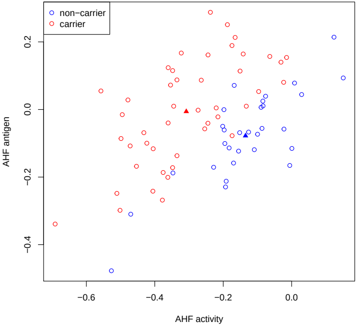
- $T^2 = [\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - (\mu_1 - \mu_2)]' \left((1/n_1 + 1/n_2)\mathbf{S}_p\right)^{-1} [\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - (\mu_1 - \mu_2)]$
- $T^2 \sim \frac{(n_1+n_2-2)p}{n_1+n_2-p-1} F_{p,n_1+n_2-p-1}$
- $\mathbf{S}_p = \frac{(n_1-1)\mathbf{S}_1+(n_2-1)\mathbf{S}_2}{n_1+n_2-2}$ is the pooled covariance matrix

Example

- Hemophilia A data. Two groups: carriers and non-carriers of a gene for Hemophilia A
- Two variables: Anti Hemophilic Factor activity (AHF-A) and AHF antigen
- Do carriers and non-carriers have the same mean vector for these variables?

	Group	AHFact	AHFanti
1	non-carrier	-0.006	-0.166
2	non-carrier	-0.170	-0.158
3	non-carrier	-0.347	-0.188
4	non-carrier	-0.089	0.006
5	non-carrier	-0.168	0.071
6	non-carrier	-0.084	0.011
7	non-carrier	-0.198	-0.000
8	non-carrier	-0.076	0.039
9	non-carrier	-0.191	-0.212
10	non-carrier	-0.109	-0.119
⋮	⋮	⋮	⋮

Group	n	AHFact	AHFanti
non-carrier	30	-0.135	-0.078
carrier	45	-0.308	-0.006



$T^2 = 40.605$, $df1 = 2$, $df2 = 72$, p-value $= 1.562e-12$

Multivariate comparison of two groups (no common $\boldsymbol{\Sigma}$)

$$H_0 : \mu_1 = \mu_2 \text{ vs } H_1 : \mu_1 \neq \mu_2$$

Assumptions:

- Both populations are multivariate normal
- $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$

Results:

- $T^2 = [\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - (\mu_1 - \mu_2)]' \left(\frac{1}{n_1}\mathbf{S}_1 + \frac{1}{n_2}\mathbf{S}_2 \right)^{-1} [\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - (\mu_1 - \mu_2)]$
- $T^2 \sim \chi^2_p$

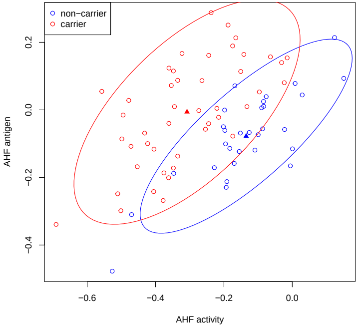
For the Hemophilia A data:

$$T^2 = 82.338, df = 2, \text{p-value} = 2.2e - 16$$

A problem

	non-carriers	
	AHFact	AHFantigen
AHFact	0.0209	0.0155
AHFantigen	0.0155	0.0179

	carriers	
	AHFact	AHFantigen
AHFact	0.0238	0.0154
AHFantigen	0.0154	0.0240



Can we assume equality of covariance matrices?

Testing equality of covariance matrices (Box M test)

$$H_0 : \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \cdots = \mathbf{\Sigma}_g \text{ vs } H_1 : \mathbf{\Sigma}_i \neq \mathbf{\Sigma}_j \text{ for some } i \neq j$$

Box M test statistic

$$M = (N - g) \ln \left(|\mathbf{S}_p| \right) - \sum_{i=1}^{g} (n_i - 1) \ln \left(|\mathbf{S}_i| \right)$$

with:

- $\mathbf{S}_p$ the pooled covariance matrix
- $\mathbf{S}_i$ covariance matrix group S_i
- N total sample size, g number of groups, n_i sample size group i

Asymptotically, the distribution of the statistic under the null:

$$X^2 = -2(1 - c) \ln (M) \approx \chi^2_{(g-1)p(p+1)/2}$$

where c is a constant for bias correction. This test is known to

- be sensitive to deviations from multivariate normality.
- have little power for small samples.
- being too liberal with large samples (rejects too often).

For the Hemophilia A data:

- $X^2 = 5.3383$, df $= 3$, p-value $= 0.1486$

Multivariate ANalysis Of Variance (MANOVA)

MANOVA is the extension of Hotelling's T^2 when there are more than two groups.

Statistical model:

$$\mathbf{x}_{\ell j} = \mu + \boldsymbol{\tau}_{\ell} + \mathbf{e}_{\ell j} = \mu_{\ell} + \mathbf{e}_{\ell j} \quad j = 1, 2, \ldots, n_{\ell} \quad \ell = 1, 2, \ldots, g \quad \mathbf{e}_{\ell j} \sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$$

Hypothesis:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_g \text{ vs } H_1 : \mu_i \neq \mu_j \text{ for some } i \neq j$$

Equivalently,

$$H_0 : \boldsymbol{\tau}_1 = \boldsymbol{\tau}_2 = \cdots = \boldsymbol{\tau}_g = \mathbf{0} \text{ vs } H_1 : \boldsymbol{\tau}_i \neq \mathbf{0} \text{ for some } i$$

- μ can be estimated by the overall sample mean vector $\bar{\mathbf{x}}$
- $\boldsymbol{\tau}$ can be estimated by the difference vectors $(\bar{\mathbf{x}}_{\ell} - \bar{\mathbf{x}})$
- $\mathbf{e}$ can be estimated by the difference vectors $(\mathbf{x}_{\ell j} - \bar{\mathbf{x}}_{\ell})$

MANOVA

- In classical univariate analysis of variance (ANOVA) the analysis consists of a decomposition of the total sum-of-squares in a between part and a within part.
- In MANOVA we have the same decomposition, but in a multivariate way.
- Matrices with sums-of-squares:

$$\mathbf{T} = \sum_{\ell=1}^{g} \sum_{j=1}^{n_\ell} (\mathbf{x}_{\ell j} - \bar{\mathbf{x}})(\mathbf{x}_{\ell j} - \bar{\mathbf{x}})'$$

$$\mathbf{B} = \sum_{\ell=1}^{g} (\bar{\mathbf{x}}_{\ell} - \bar{\mathbf{x}})(\bar{\mathbf{x}}_{\ell} - \bar{\mathbf{x}})'$$

$$\mathbf{W} = \sum_{\ell=1}^{g} \sum_{j=1}^{n_\ell} (\mathbf{x}_{\ell j} - \bar{\mathbf{x}}_{\ell})(\mathbf{x}_{\ell j} - \bar{\mathbf{x}}_{\ell})'$$

- and it holds that

$$\mathbf{T}_{p \times p} = \mathbf{B}_{p \times p} + \mathbf{W}_{p \times p}$$

MANOVA table

Source	Sums-of-Squares	DF
Treatment	$\mathbf{B}$	$g-1$
Residual	$\mathbf{W}$	$\sum_{\ell=1}^{g} n_\ell - g$
Total	$\mathbf{T}$	$\sum_{\ell=1}^{g} n_\ell - 1$

To test the null, we use Wilks' lambda

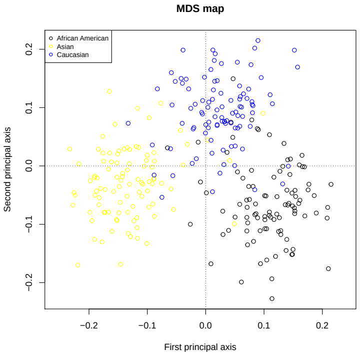
$$\Lambda = \frac{|\mathbf{W}|}{|\mathbf{B}+\mathbf{W}|}$$

For large samples

$$-\left(n-1-\frac{p+g}{2}\right)\ln(\Lambda) \sim \chi^2_{p(g-1)}$$

Alternative statistics, such as Pillai's trace or Roy's largest root are often used, and equivalent to Wilks' Λ for large samples.

MANOVA Example: NIST data



	Wilks	pvalue
Axes 1-2	0.1087	0.0000
Axes 3-10	0.9342	0.2509

Bibliography

- Johnson & Wichern, (2002) Applied Multivariate Statistical Analysis, Chapters 4 and 5, 5th edition, Prentice Hall.